2D Pose-based Real-time Human Action Recognition with Occlusion-handling

Federico Angelini, *Student Member, IEEE,* Zeyu Fu, *Member, IEEE,* Yang Long, *Senior Member, IEEE,* Ling Shao, *Senior Member, IEEE*, and Syed Mohsen Naqvi, *Senior Member, IEEE*

***Abstract*—Human Action Recognition (HAR) for CCTV-oriented applications is still a challenging problem. Real-world scenarios HAR implementations is difficult because of the gap between Deep Learning data requirements and what the CCTV-based frameworks can offer in terms of data recording equipments. We propose to reduce this gap by exploiting human poses provided by the OpenPose, which has been already proven to be an effective detector in CCTV-like recordings for tracking applications. Therefore, in this work, we first propose ActionXPose: a novel 2D pose-based approach for pose-level HAR. ActionXPose extracts low- and high-level features from body poses which are provided to a Long Short-Term Memory Neural Network and a 1D Convolutional Neural Network for the classification. We also provide a new dataset, named ISLD, for realistic pose-level HAR in a CCTV-like environment, recorded in the Intelligent Sensing Lab. ActionXPose is extensively tested on ISLD under multiple experimental settings, e.g. Dataset Augmentation and Cross-Dataset setting, as well as revising other existing datasets for HAR. ActionXPose achieves state-of-the-art performance in terms of accuracy, very high robustness to occlusions and missing data, and promising results for practical implementation in real-world applications.**

***Index Terms*—Pose, LSTM, CNN, ISLD, CCTV**

I. Introduction

HUMAN Action Recognition (HAR) is one of the most challenging problems for Artificial Intelligence (AI) [1]. It consists of training an AI model to recognise a class of human actions. Depending on the applications, input data can be very different, such as RGB videos, infrared data, time-of-flight-based or structured-light-based data (depth data) [2].

HAR is a crucial task in many applications, such as cybernetics, human-machine interaction, automated assisted living systems, surveillance, autonomous vehicles, gaming and sports analysis. In this work, the focus is on surveillance and CCTV-like data. We chose to focus on this area because artificial intelligence for surveillance-related HAR is yet to make an impact on practical applications. Therefore, this work presents ActionXPose, a real-time body pose-based method for HAR, which is robust to occlusions and multi-viewpoint changes. In Fig. 1, an example of human 2D poses provided by OpenPose [3] is shown.

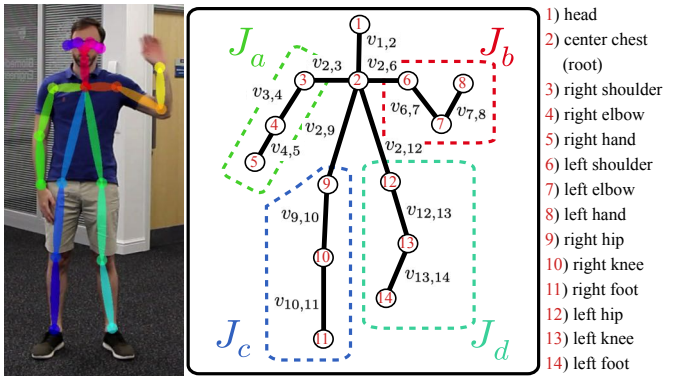


Fig. 1. The human pose is a set of 2D landmarks J, which represents a non-redundant description of the target posture. (Right) Examples of a human pose. (Left) Proposed pose nomenclature, for pose landmarks $J = \{1, \ldots, 14\}$, link vectors v_{j_1,j_2}, for $j_1, j_2 \in J$, and considered landmark subsets, i.e. J_a, J_b, J_c and J_d.

This work is based on the detection framework shown in Fig. 2, which we have already explored for tracking-related problems [4]–[8]. When multi-target behaviour analysis based on CCTV-like recordings is required, the first step, generally, consists of the detection of human targets by using RGB data. This step requires significant computational effort. Subsequently, tracking is needed to retrieve targets identities. The tracking processing relies primarily on detected bounding boxes or body landmarks coordinates rather than on the RGB data. Thus, this step can be performed quite efficiently. Ultimately, each tracked target data can be further processed for HAR. However, this step can be computationally expensive if bounding boxes RGB data is considered. As opposite, only considering bounding boxes coordinates could be very computationally efficient. However, in this case, each target data is reduced to a centroid point, and no posture-related HAR can be performed with such limited data. Therefore, the aim of this work is to find a trade-off between computational cost and informativeness by using the OpenPose detector for HAR already exploited for the tracking phase. Therefore, RGB data is intentionally neglected from the HAR processing to compensate for the computational effort already spent during the detection phase. This strategy enables to focus on each targets postures and movements, preserving and exploiting the tracking results, to reduce the computational cost. However, pose detectors are prone to false landmarks detections due

F. Angelini, Z. Fu and S. M. Naqvi are with the Intelligent Sensing and Communications Research Group, School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK (e-mails: f.angelini2@ncl.ac.uk; z.fu2@ncl.ac.uk; mohsen.naqvi@ncl.ac.uk).

Y. Long is with Department of Computer Science, Durham University, Durham DH1 3LE, UK (e-mail: yang.long@durham.ac.uk).

L. Shao is with Inception Institute of Artificial Intelligence, Abu Dhabi, UAE (e-mail: ling.shao@ieee.org).

Corresponding e-mail address: Mohsen.Naqvi@ncl.ac.uk

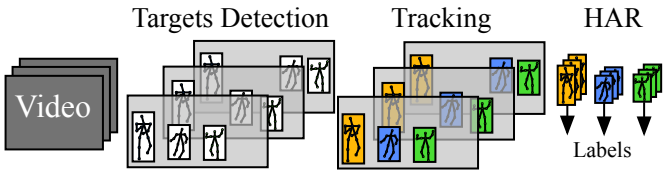


Fig. 2. Overall HAR framework. CCTV-like videos are pre-processed with a human pose detector, such as OpenPose, to estimate the positions of the targets (bounding box) and their body limb positions (landmarks). Subsequently, tracking algorithms provide targets identities, allowing consistent grouping of detected targets data for HAR. Finally, for each detected target, the action label is estimated by using the tracked data.

to occlusions or cluttered RGB data. This problem has been already explored and documented in our recent work for tracking applications [8]. Thus, this work focuses on proposing strategies for pose-based HAR which can be robust to occlusions and missing data, for CCTV-based applications.

We considered CCTV-like recordings, such that:

1) The camera viewpoint is far enough to capture most of the target bodies in absence of occlusions;
2) The frame resolution and compression allow reasonable OpenPose detection.

Moreover, this work proposes a new dataset, namely the Intelligent Sensing Lab Dataset (ISLD), which is specifically designed for pose-based HAR in a CCTV-like recording environment. To further extend the learning abilities of ActionXPose, additional existing datasets are also explored. To evaluate ActionXPose performance, we performed single-dataset and cross-dataset testing, dataset augmentation testing, ablation study, occlusions study and video quality study. Results show that ActionXPose outperforms existing methods in most of the tests and shows greater robustness to occlusions and missing data.

In conclusion, our main contribution is threefold:

1) A new, real-time algorithm for posed-based HAR, robust to occlusions and missing data;
2) An extensive exploration on dataset augmentation using existing datasets to improve the proposed model is proposed;
3) A new dataset, named ISLD, for pose-based HAR.

The rest of this paper is organised as follows: in Section II, related works are presented. Section III includes technical details about ActionXPose, regarding low-level features extraction, high-level features computation and occlusion-handling strategies definition. In Section IV, several experiments are presented to extensively evaluate the performance of ActionXPose. Finally, in Section V, conclusions are drawn and future work is proposed.

II. Related Work

The state-of-the-art HAR algorithms can be divided into three main categories [1]:

1) Hand-crafted features based approaches: in this category, human insights are strongly used to solve HAR problems, while machine abilities are not fully exploited and are often limited to conventional machine learning tasks. For example, notable approaches rely on background subtraction methods [9], 3D Histogram of Oriented Gradients (3D-HOG) [10] and Local Binary Patterns (LPB) [2], [11].

2) Deep-learning based approaches: in this category, Convolutional Neural Networks (CNNs), generative models, 3D-CNNs and Recurrent Neural Networks (RNNs) [12] are used to explore data without any or with very limited human insights.

3) Hybrid approaches: the algorithms in this category attempt to combine the most promising results from both the above mentioned hand-crafted and deep learning based approaches, providing a useful trade-off between them [1].

According to the above taxonomy, the method presented in this work, ActionXPose, is a fully-fledged hybrid approach for pose-level HAR. Indeed, ActionXPose is based on human poses extracted by a deep learning based detector, i.e. OpenPose. Then, poses are processed to extract low and high-level features represented as time-sequences. This process is supported by Self-Organizing Map (SOM) networks [13] and commonly used tools such as the Principal Component Analysis (PCA). Produced time-sequences are further processed by the MLSTM-FCN network [14]. This state-of-the-art network combines a Long Short-Term Memory (LSTM) [15] block with a *squeeze and excitation* block, to better consider both time dependencies and mutual relationships between sequences.

While human poses consist of 2D data, the human skeletons provided by the popular structured-infrared-light Kinect camera consist of 3D data. Skeleton-based approaches are strongly related to the proposed method. However, despite the significant advantages provided by the 3D data [16], for example in terms of viewpoints generalization, Kinect presents numerous limitations. For example, it does not work well in outdoor environments and has a minimal working range (up to 5-6 meters), which limits its ability to be implemented in surveillance scenarios [17]. Nevertheless, the skeleton based literature can still provide several cues on how to process the 2D poses. However, it is worth emphasizing that, for example, the 2D poses do not allow any rotation in the 3D space. Thus, many skeleton-based algorithms, which are explicitly or implicitly based on the advantageous properties of the 3D space, are limited in the 2D case. Furthermore, no depth data, i.e. Kinect points cloud, is available in the 2D case. Despite these problems, the basic body landmarks processing can still be borrowed from the 3D skeleton approaches. For example, landmarks normalization is commonly used to remove target size and location dependency. At a feature-level, it is common to consider the mutual Euclidean distance between landmarks. In fact, in [18], authors proposed to extract Local Occupancy Pattern (LOP) features from depth data and Invariant Features from skeleton data, where landmarks normalization was mainly used. Similarly, in [19], authors suggested to implement Hidden Markov Models (HMMs) to process skeleton data where normalization was performed as a pre-processing step.

Motivated by the limitations of Kinect, researchers started developing new techniques that can provide similar output data using a simple RGB camera. Therefore, highly promising body

pose detectors have been published in the last few years, such as DeeperCut [20] and OpenPose [3]. In particular, OpenPose achieves the best performance, and opens a new research direction for HAR. For example, in [21], Yan et al. have implemented graph convolutional networks to process pose data provided by OpenPose, which achieves promising results. This approach considers temporal information alongside spatial information. However, this method requires a fixed number of frames for each action sample in order to build the action graph, which may affect system flexibility. In contrast, our work focuses on exploiting LSTM based networks. Thus, our work is specifically designed to deal with multivariate temporal sequences, with no restrictions on the number of time steps, allowing full flexibility with respect to space and time.

Regarding the classification step, RNN and LSTM have also been exploited in many recent papers. In [22], the authors focused on the 3D skeleton-based action recognition using LSTM networks. Their major focus was on implementing trusting gates on the LSTM architecture to allow better action representation. In [23], a hierarchical RNN approach was implemented to focus on several 3D skeleton subparts, in order to better discriminate which body sub-part is related to the performed action. In [24], authors focused on new gate strategies for LSTM, in order to emphasize salient motion from learning data. In [25], a new regularization term for LSTM was used in order to learn co-occurrences within skeleton data.

Most of the above mentioned studies focused on modifying the network architectures to process the raw 3D skeleton data more effectively. In contrast, our proposed work aims to extract more meaningful and robust input features based on the 2D body landmark coordinates, to be provided as input for a predefined LSTM architecture.

Other approaches relevant to this work for performance comparison are [10], [26]–[34]. In these works, the learning sources are mainly RGB raw data, motion history videos and body silhouettes. Compared with human poses, these data are generally heavier, redundant and cluttered. Data abundance can provide more effectiveness for HAR, at the cost of more processing time. Moreover, in those works, no explicit link with human tracking was provided. In contrast, the proposed approach has a low processing cost, and it is specifically designed for tracking-based frameworks.

III. Methodology

A. Problem's Statement and Notations

Let $\mathbb{D} = \{s_i, l_i, w_i\}_{i=1}^{N}$ be the action dataset containing N samples. s_i represents the i-th subject video sample. $l_i \in \mathcal{L}$ represents the i-th action label, where $\mathcal{L}$ is the set of target actions and $w_i \in \mathcal{W}$ represents the i-th viewpoint label, where $\mathcal{W}$ is the set of considered viewpoints. Let $\mathbb{T} \subset \mathbb{D}$ be the chosen training subset and $\mathbb{T}^* = \mathbb{D} \backslash \mathbb{T}$ be the testing subset.

The pose detector provides a root-centred graph in the form of 2D coordinates, such that

$$p_i(t) = \{(x_j(t), y_j(t))_i\}_{j \in J} \quad t \in \{1, \dots, T_i\} \tag{1}$$

where T_i represents the time length of the i-th video sample and J is the landmarks set defined by the pose detector mapping. In this work, $J = \{1, \dots, 14\}$. Let v_{j_1,j_2} be the link vector between body landmarks $j_1, j_2 \in J$, as defined in Fig. 1.

ActionXPose aims to exploit $p_i(t)$, as well as generate additional and more robust time sequences, to train a recurrent neural network using $\mathbb{T}$ and to predict action labels in $\mathbb{T}^*$.

B. Baseline Methods

In this section, three baseline methods are defined, which borrow techniques from the 3D skeleton-based HAR field.

Sequence $p_i(t)$ provided by OpenPose are location and body size dependent [19]. Thus, translation and scaling are required in order to normalize data across different samples. Specifically, the location dependency problem can be addressed by transforming $p_i(t)$ from the *absolute* to the *root-centred* coordinate reference system. The transformed pose coordinates are defined as:

$$(\bar{x}_j, \bar{y}_j)_i = (x_j, y_j)_i - (x_2, y_2)_i \quad \forall j \in J, \tag{2}$$

where the $\bar{(.)}$ operator denotes the centring transformation and where the dependence of t has been conveniently omitted. Thus,

$$\bar{p}_i = \{(\bar{x}_j, \bar{y}_j)_i\}_{j \in J}, \tag{3}$$

is the set of root-centred coordinates defined by (2).

Furthermore, we denote the scaling with the $\bar{\bar{(.)}}$ operator as follows:

$$\bar{\bar{p}}_i = \{(\bar{\bar{x}}_j, \bar{\bar{y}}_j)_i\}_{j \in J} \tag{4}$$

where $\bar{\bar{p}}_i$ is obtained by scaling $\bar{p}_i(t)$ coordinates by using the following constraint

$$\bar{\bar{v}}_{j_1,j_2} = \frac{\bar{v}_{j_1,j_2}}{||\bar{v}_{2,9}||_2} \quad \forall j_1, j_2 \in J \tag{5}$$

where $\bar{v}_{2,9}$ is the vector link between the root and the right hip landmarks. Due to (2) and (5), the target position and the size information are discarded.

According to the proposed definition for $p_i(t)$, these sequences mostly contain *spatial* information about the motion. To obtain *temporal* information, $p'_i(t) = p_i(t+1) - p_i(t)$ can be defined. For the rest of the paper, we will use $p_i(t)$ to denote the transformed poses in (4). Therefore, inspired by existing literature, we have selected three baseline methods, which are closely related the proposed problem.

1) [19] + [14]: this consists of a simple learning step based on normalized OpenPose coordinates sequences $p_i(t)$ and $p'(t)$. A Multivariate LSTM-FCN architecture with a *time-based attention mechanism* (MLSTM-FCN) [14] is used for the classification step. This algorithm takes as input the coordinate sequences obtained from training data $\mathbb{T}$, including action labels l_i, to train a supervised classification model to be tested on $\mathbb{T}^*$.

2) [18] + [14]: this consists of computing mutual OpenPose landmarks distances [18]. In this case, the classification step is also performed by using the MLSMT-FCN architecture.

3) [18] + [19] + [14]: this consists of a hybrid approach merging the previous two methods. Thus, $p_i(t)$, $p'(t)$ and mutual distances between landmarks are considered. The classification step is again performed by using the MLSMT-FCN architecture.

Formally, $p_i(t)$ and $p'_i(t)$ represent *low-level* features, which can be provided to the MLSTM-FCN for classification. In particular, by using $p_i(t)$ and $p'_i(t)$, the MLSTM-FCN performs landmark-based attention due to its architecture. However, such levels of detail can be confusing in some cases, due to high intra-class similarities or within-class variations. Moreover, when some landmarks are persistently missing due to occlusions, the corresponding sequences will be completely lost, compromising the robustness against unexpected occlusions. For example, for the baseline methods in III-B2 and III-B3, one single persistent missing landmark $(x_j, y_j)_i$ not only neglects two x_j and y_j sequences, but also compromises the calculation of mutual distances. Therefore, the next sections provide novel high-level features, that are designed to be robust to missing data, and additional occlusion-handling methods, providing an effective solution to this problem.

C. Defining Poses Libraries

The main goal of this section is to exploit training data $\mathbb{T}$ to learn general poses that best represent each action, from each viewpoint. In other words, the output of this step is a *pose library*.

Since root coordinates in $p_i(t)$ were set to zero in the previous section, let $u_i(t) = (x_1, y_1, x_3, y_3, \ldots, x_J, y_J) \in \mathbb{R}^{2J-2}$ be the vector obtained by unrolling $p_i(t)$ and skipping the root coordinates $(x_2, y_2)_i$. The unsupervised clustering method Self-Organizing Map (SOM) [13] is used in a semi-supervised fashion, to explore natural clusters in $\mathbb{R}^{2J-2}$. Since the SOM algorithm expects no missing data, in this stage, we exploited body left/right symmetry for dealing with possible persistent occlusions occurred in training data, mostly due to target self-occlusions. In particular, we estimated the persistently missing landmarks by mirroring available data. For example, if the left-shoulder was missing, data was filled with the transformed right-shoulder obtained by mirroring it with respect to the root landmark.

Additionally, SOM requires a cluster topology to be defined. Since prior-information about the distribution of pose data is not provided, the homogeneous topology, $[q, \ldots, q] \in \mathbb{R}^m$, is set for a given integer q and a given space dimension m. This choice forces the SOM architecture to have q^m neurons linked each other with a homogenous rectangular topology, defining q^m clusters. Since the SOM computational time is affected by either the number of considered vectors $u_i(t)$, the q topology parameter and the space dimension $2J-2$, a trade-off between these parameters is required.

To solve this problem, let $\tilde{u}_i(t)$ be the vector containing the first m principal components of $u_i(t)$ obtained through the PCA, i.e. $\tilde{u}_i(t) \in \mathbb{R}^m$. Our simulations suggest that the best values for m and q are $m = 3$ and $q = 4$, which balance the SOM computational cost while producing a reasonable number of prototypes. In Fig. 3, comparisons of SOM computational times are provided. Therefore, the whole process can be summarised as follows

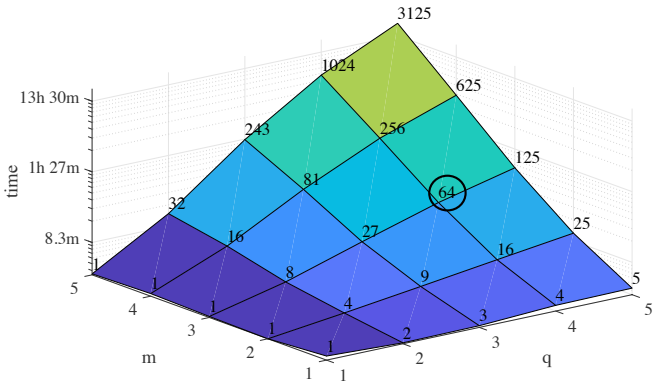


Fig. 3. SOM time computation comparison. Different values for m and q have been set for the SOM computation. Reported times refer to entire library creation process for the MPOSE dataset. In the graph, q^m values are also reported. As shown, $q^m = 4^3 = 64$ is chosen as trade-off between computation time and number of prototypes in the libraries.

$$\mathbb{R}^{2J-2} \xrightarrow{\text{PCA}} \mathbb{R}^m \xrightarrow{\text{SOM}} q^m \text{ clusters} \tag{6}$$

Thus, for a fixed action label l and a fixed point of view w, the SOM is trained over

$$\{\tilde{u}_i(t) \mid l_i = l, w_i = w\} \subset \mathbb{T}. \tag{7}$$

This provides an additional cluster label k_i for each training pose $\tilde{u}_i(t)$, as follows:

$$\{\tilde{u}_i(t) \mid l_i = l, w_i = w\} \xrightarrow{\text{SOM}} \{\tilde{u}_i(t) \mid l_i = l, w_i = w, k_i = k\}$$
$$\forall\, l \in \mathcal{L}, \quad \forall\, w \in \mathcal{W}, \quad k \in \{1, \ldots, q^m\}. \tag{8}$$

Thus, q^m pose prototypes are defined by averaging cluster labels k_i as follows:

$$U_{l,w,k} = \frac{1}{n_k} \sum_{i,t} \{u_i(t) \mid l_i = l, w_i = w, k_i = k\}$$
$$\forall\, l \in \mathcal{L}, \quad \forall\, w \in \mathcal{W}, \quad k \in \{1, \ldots, q^m\}, \tag{9}$$

where n_k represents the number of poses within cluster k. In conclusion of this step, the libraries of prototypes are collected from training data as follows:

$$V_l = \left\{ \{U_{l,1,k}\}_{k=1}^{q^m}, \ldots, \{U_{l,|\mathcal{W}|,k}\}_{k=1}^{q^m} \right\} \quad \forall\, l \in \mathcal{L}. \tag{10}$$

Thus, V_l contains pose prototypes in the form of points in a multidimensional space $\mathbb{R}^{2J-2}$, which are able to cover all variation of considered viewpoints. For a visual example of the V_l set, see Fig. 4.

Libraries for the temporal information can be similarly defined as follows:

$$S_l = \left\{ \{U'_{l,1,k}\}_{k=1}^{q^m}, \ldots, \{U'_{l,\mathcal{W},k}\}_{k=1}^{q^m} \right\} \quad \forall\, l \in \mathcal{L}, \tag{11}$$

where $U'_{l,w,k}$ represents prototypes obtained by clustering temporal vectors $u'_i(t)$.

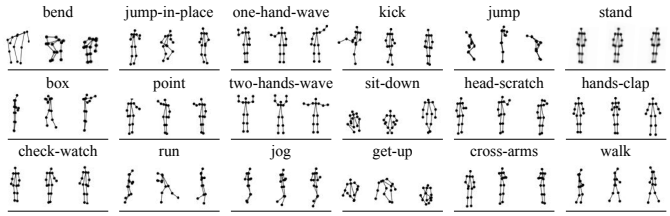


Fig. 4. Spatial library prototype examples. Three prototypes are randomly selected from V_l, for all $l \in \mathcal{L}$.

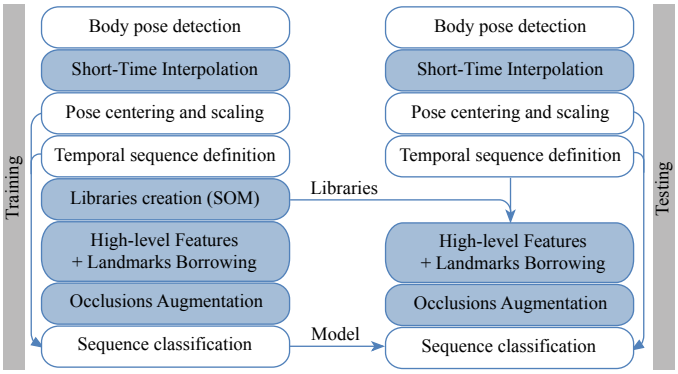


Fig. 5. Proposed training and testing ActionXPose pipelines. The occlusions-handling steps are depicted in blue.

D. Strategies for Occlusion-Handling

Occlusions, self-occlusions or ambiguous RGB data can affect OpenPose performance, resulting in *persistent* or *short-time* missing data. In this section, we propose four complementary strategies to deal with these problems.

1) High-level Features: In this section, we address the problem of persistent occlusions. A persistent occlusion occur when one or more landmarks are missing for the *entire* sequence. To address this problem, the idea is to exploit the Spatio-temporal libraries V_l and S_l for $l \in \mathcal{L}$ defined in the previous section, to generate high-level features in the form of time sequences. Inspired by [23], since different body parts carry different information, the idea is to exploit full-body and local-limb attention.

Given $J = \{1, \ldots, 14\}$, let J_a, J_b, J_c, J_d be the landmark subsets as defined in Fig. 1; namely,

$$J_a = \{3,4,5\} \subset J \qquad J_b = \{6,7,8\} \subset J$$
$$J_c = \{9,10,11\} \subset J \qquad J_d = \{12,13,14\} \subset J. \tag{12}$$

Let $d_{J_*}(p_i(t), v)$ be the average distance between the generic pose $p_i(t)$ and the generic prototypes $v \in V_l$, computed for landmarks J_*, where J_* represents either J_a, J_b, J_c, J_d or J, and it is defined as follows:

$$d_{J_*}(p_i(t), v) = \frac{1}{|\bar{J}_*|} \sum_{j \in \bar{J}_*} \|(x_j, y_j)_i - (x_j, y_j)_v\|_2 , \tag{13}$$

where $(x_j, y_j)_v$ are the j-th landmark coordinates of v and $\bar{J}_*$ represents either J_a, J_b, J_c, J_d or J, where missing coordinates are excluded. Therefore, given a library of prototypes V_l for action l, we can define the *embedding sequence* as follows:

$$D_{V_l,J_*}(t) = \min_{v \in V_l} d_{J_*}(p_i(t), v), \tag{14}$$

where it is clearly shown that D_{V_l,J_*} depends on time.

Given a set of actions $\mathcal{L}$ and the set of landmarks in (12), the meaningful sequences that can be extracted from $p_i(t)$ are defined as follows:

$$Seq_i(V_l) = \{D_{V_l,J}(t), D_{V_l,J_a}(t), \ldots$$
$$\ldots, D_{V_l,J_b}(t), D_{V_l,J_c}(t), D_{V_l,J_d}(t)\} \quad \forall l \in \mathcal{L}, \tag{15}$$

Similarly, sequences for temporal information can be embedded as follows:

$$Seq_i(S_l) = \{D_{S_l,J}(t), D_{S_l,J_a}(t), \ldots$$
$$\ldots, D_{S_l,J_b}(t), D_{S_l,J_c}(t), D_{S_l,J_d}(t)\} \quad \forall l \in \mathcal{L} \tag{16}$$

where we simply replace the term V_l in (15) with S_l. This leads to two sets of sequences, $Seq_i(V_l)$ and $Seq_i(S_l)$, for all $l \in \mathcal{L}$.

It is worth mentioning that (13) allows the embedding to run and provide numerical results, even in presence of persistent missing data, due to the presence of $\bar{J}_*$. In other words, when missing data occurs, the sequences provided by (15) and (16) are lost only when *all* landmarks in the selected landmarks set are missing. In all other cases, i.e. at least one landmark is available for the selected landmarks set, the distance computation is successfully performed and the corresponding sequence is obtained.

We will prove in Section IV-F that this approach not only preserves the classification integrity in case of occlusions, but also improves baseline performance.

2) Landmark Borrowing: in this section, we provide a strategy to improve low-level features in case of persistent occlusions. Equation (13) is based on $\bar{J}_*$, which only contains non missing landmarks. Thus, the resulting sequences in equation (15) and (16) are well-defined even in presence of missing landmarks. However, low-level sequences $p_i(t)$ and $p'(t)$ might still show missing values when an occlusion occur. To solve this problem, we propose to further exploit equations (13) and (14) to fill the missing values in $p_i(t)$ by using the knowledge contained in the pose libraries. Therefore, for a given time step t,

$$v^\dagger(t) = \arg\min_{v \in V_l, l \in \mathcal{L}} d_J(p_i(t), v), \tag{17}$$

is the *closest prototype* to the pose $p_i(t)$. Thus, we can exploit $v^\dagger(t)$ to borrow the coordinates of the missing landmarks in $p_i(t)$. Subsequently, $p'(t)$ is computed according to the usual definition. This strategy ensures to fill missing data in the low-level sequences. Moreover, no extra cost is required to perform

Fig. 6. Examples from the proposed ISLD dataset. Human actions of 10 subjects are recorded using a static camera, from different viewpoints.

equation (17), since the calculation can be embedded within the one required by equation (14).

3) Short-time Interpolation: in this section, we address the problem of short-time occlusion. This problem occur when, within the considered sequence, landmarks coordinates are missing for only few frames. This case is much easier than the one considered in Section III-D1, where persistent occlusions where addressed instead. Although Kalman filter [35] can be applied to the detected landmarks, further processing is required to ensure that the Gaussian property holds for the considered data. Thus, our strategy to deal with short-time missing data consists of interpolating available data, exploiting temporal consistency. In our formulations, we define $x(t)$ and $y(t)$ as the landmark coordinate with respect to time t, where some entries are occasionally missing (short-time), such that:

$$x : A \longrightarrow \mathbb{R}, \quad y : A \longrightarrow \mathbb{R}, \quad A \subset \{1, \ldots, T\}, \tag{18}$$

where A represents the set of frames when the landmark is detected. Then, we define missing values for $t^* \in \{1, \ldots, T\} \setminus A$ as the *nearest-neighbour*:

$$x(t^*) = x\left(\hat{t}\right), \quad y(t^*) = y\left(\hat{t}\right), \quad \text{s.t.} \quad \hat{t} = \underset{t \in A}{\arg\min} \|t^* - t\|_2. \tag{19}$$

Given the simplicity of this solutions, we implemented it at all stages of our work, including for the baseline methods, as a simple and reasonable quick solution for short-time missing data.

4) Occlusions Augmentation: as a final strategy for occlusion-handling, synthetically occluded sequences are added in the training phase. Specifically, training sample can be persistently occluded by randomly removing some landmarks according to a binary Bernoulli distribution $\mathcal{B}(p)$ where $p = 0.5$. This strategy has been implemented right after landmark detection. To preserve the integrity of the system, landmarks 2 and 9 have been not occluded to allow equation (5) to be well defined. This strategy aims to train resulting network with data that present random occlusions, enabling the network to learn a more general representation. This strategy can provides additional robustness. In fact, despite the presence of the other strategies, it is crucial to effectively learn how the low and high-level features might change when different occlusions occur.

E. Classification Step

For fair comparisons with the proposed baseline methods, MSTLM-FCN is again used for the classification step. Depending on the input features, the classification step can focus on different motion aspects. Specifically, in this work, we define three sets of sequences, by using the proposed low-level and high-level features, as follows:

1) Spatial-attention sequences: these are formed by combining $p_i(t)$, $p'_i(t)$ and $Seq_i(V_l)$.

2) Temporal-attention sequences: these are formed by combining $p_i(t)$, $p'_i(t)$ and $Seq_i(S_l)$.

3) Spatio-temporal-attention sequences: these are formed by combining $p_i(t)$, $p'_i(t)$, $Seq_i(V_l)$ and $Seq_i(S_l)$.

For an overview of the ActionXPose processing, in Fig. 5 the general pipeline of the proposed algorithm is provided.

IV. Experiments

A. ISLD Dataset

In this work, we propose a new, realistic dataset for the pose-based HAR, named ISLD. This dataset was recorded within our Intelligent Sensing Lab. Single-target CCTV-like clips, according to 18 predefined posture-related action classes, were collected. Participants were free to perform the actions according to their understanding of the class labels and no example clips were provided. Recording viewpoints were predefined, to ensure that enough viewpoints were covered. Specifically, samples were recorded from up to 5 different viewpoints, namely *front, front-left, front-right, left* and *right*. The 18 proposed actions, performed multiple times by 10 actors, were recorded with a static RGB camera. Overall, ISLD contains 907 different time windows. For each time window, only one target is visible, performing a single action. 10 examples from the ISLD dataset are shown in Fig. 6.

For the purpose of this work, we pre-processed ISLD samples with OpenPose to extract human poses. To increase the number of samples for the classification requirements, data augmentation was performed. In fact, deep learning methods usually require a great amount of data to perform well. In fields such as image recognition, *cropping* or *rotating* images are common practice to augment dataset samples in order to meet deep learning algorithms conditions [36]. In speech recognition, it is also common to add noise to training samples for the same purpose [37]. In this work, two methods for augmenting training data were used.

The first method is named *pose-flipping*, which consists of flipping poses along the vertical axis passing through the root landmark. This causes that the performed action looks mirrored, exploiting the left/right body symmetry. In Fig. 7, viewpoint composition rates for the ISLD dataset are provided, showing that pose-flipping balances the left/right viewpoint rates.

The second method for data augmentation is named *pose-noising*, which consists of adding Gaussian noise to the landmark coordinates, i.e. $\mathcal{N}(0, \sigma^2)$ with 0 mean and σ standard deviation. In this work, $\sigma = 0.2$ is empirically chosen for all experiments unless otherwise specified. Specifically, let z be the number of times that training data are used to create additional noisy samples. Thus, if $z = 0$, no noisy samples were created. If $z = 1$, all training samples were used *once* to create noisy samples.

In conclusion, after applying the proposed data augmentation, the ISLD dataset consists of up to 5598 samples, as shown in Table I. Fig. 7 shows that pose-flipping not only doubles the available data, but also balances left/right viewpoint rates.

B. Experimental Settings

1) Traditional Setting: this setting is fully based on the ISLD dataset. ActionXPose was trained on actors $[1, 2, 3, 4]$, validated on actors $[5, 6, 7]$ and tested on actors $[8, 9, 10]$. Regarding hyper-parameters, $\sigma = 0.2$ and $z = 1$ for augmenting training data.

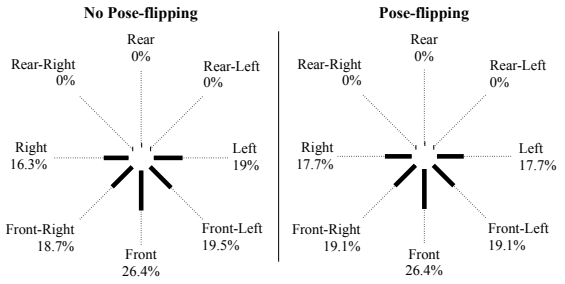


Fig. 7. ISLD viewpoints composition rates in the cases of no pose-flipping and pose-flipping data augmentation. Pose-flipping is useful for doubling the number of samples and balancing viewpoints composition.

TABLE I
ISLD ACTION COMPOSITION. THE NUMBER OF SAMPLES FOR EACH ACTION ARE PROVIDED IN THE DEFAULT CASE (NO DATA AUGMENTATION) AND AFTER APPLYING POSE-FLIPPING AND POSE-NOISING.

	D. Aug.			D. Aug.	
Label	**No**	**Yes**	**Label**	**No**	**Yes**
bend	40	240	jump	40	240
box	80	480	kick	42	252
check-watch	40	240	jump-in-place	48	288
cross-arms	40	240	point	40	240
get-up	40	240	run	40	240
hand-clap	40	240	head-scratch	40	240
one-hand-wave	40	240	sit-down	40	240
two-hands-wave	40	240	stand	350	1050
jog	36	216	walk	72	432

2) Dataset Augmentation Setting: in this setting, ISLD training data were considered alongside additional datasets to better leverage the deep learning generalization ability. Since the *stand* action is already well covered by ISLD, no further data augmentation was required for this action. However, other classes are not so well represented and additional data can be helpful. Because data collection for HAR is often expensive and time-consuming, we revised already available datasets, starting with the popular UCF101 [38] and HDMB51 [39] datasets. For these two datasets, most of the video samples were collected from YouTube and movies. The camera was often too close to the target, capturing only the target's face or hands. Moreover, most samples in these datasets show low-resolution, unlabelled, multiple-target frames where the subjects perform different actions. Furthermore, most of the actions are strongly related to the context rather than to the human posture. Last but not least, as shown in Fig. 8-(Top), if OpenPose is used to pre-process these datasets, the overall performance is too low to be a reliable source for the proposed 2D pose-based HAR. Fig. 8-(Bottom), shows the OpenPose detection rate for UCF101 and HDMB51, supporting these conclusions.

In contrast to UCF101 and HDMB51, CCTV-like recordings often show full-body targets, where OpenPose works well. Fig. 8-(b) also shows the performance on a famous dataset for tracking in public environments, i.e. MOT16 [40], and other traditional datasets, i.e. KTH [41], IXMAS [42], Weizmann [43] and i3DPost [44]. We found that OpenPose performs considerably better on these traditional datasets. Moreover, these datasets include fully-labelled single-target clips, which

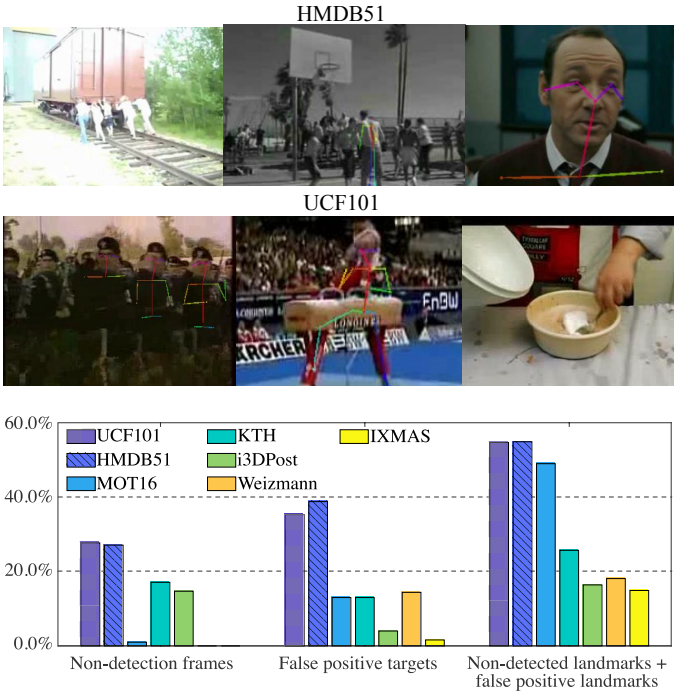


Fig. 8. (Top) Screenshots from HMDB51 and UCF101 datasets processed by OpenPose. The pictures show false negatives, missing landmarks, false positives and very limited views of the target body. (Bottom) OpenPose performance on different datasets. *Non-detection frames* measures the percentage of frames in which detections do no occur. *False positive targets* measures the percentage of the root landmark confidence below the threshold 0.5. Considering those targets with the root landmark confidence higher than 0.5, *Non-detected landmarks + false positive landmarks* considers the percentage of non detected landmarks (confidence = 0), plus false positive landmarks ($0 <$ confidence < 0.5).

simplifies our work removing the requirement of the tracking pre-processing step. Motivated by these considerations, we collected additional training data from the Weizmann, i3DPost, KTH and IXMAS datasets, defining the four-in-one MPOSE dataset by merging them all together. Moreover, the MPOSE class labels were selected to be consistent with the ISLD's labels. Overall, MPOSE contains 4160 single-target video clips, distributed over 17 action classes (the *stand* action is excluded) and performed by 53 human actors. Table II reports the MPOSE action composition for the default case (no data augmentation) and the pose-flipping/pose-nosing case. Pose-noising has the potential to indefinitely raise the total number of samples. However, for the MPOSE dataset, we experimentally found that $z = 2$ is the best pose-noising parameter. It is worth noticing that, when data augmentation is applied, MPOSE contains a significantly higher number of samples than UCF101 and HMDB51, which contain 13320 and 6766 samples, respectively. In terms of viewpoints, Fig. 9 shows the viewpoint composition and the effect of pose-flipping in balancing left/right viewpoints rate.

3) Cross-Dataset Setting: in this setting, the MPOSE dataset was used for training and validation, while the whole ISLD dataset was used for testing. Therefore, the purpose of this test was to measure ActionXPose cross-dataset performance. We remark that, in this setting, since MPOSE does not contain data for the action *stand*, we neglected this action

TABLE II
MPOSE ACTION COMPOSITION. THE NUMBER OF SAMPLES FOR EACH ACTION ARE PROVIDED IN THE DEFAULT CASE (NO DATA AUGMENTATION) AND AFTER APPLYING POSE-FLIPPING AND POSE-NOISING.

	D. Aug.			D. Aug.	
Label	**No**	**Yes**	**Label**	**No**	**Yes**
bend	193	1158	jump	73	438
box	517	3102	kick	120	720
check-watch	120	720	jump-in-place	73	438
cross-arms	120	720	point	120	720
get-up	120	720	run	474	2844
hand-clap	396	2376	head-scratch	120	720
one-hand-wave	193	1158	sit-down	120	720
two-hands-wave	407	2442	stand	0	0
jog	400	2400	walk	594	3564

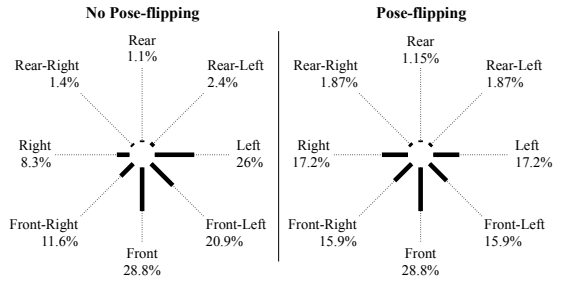


Fig. 9. MPOSE viewpoint composition in both cases of no pose-flipping and pose-flipping.

in ISLD as well.

C. Implementation

Simulations were conducted on Ubuntu 16.04 running on a Dell Inspiron 15 5000 with four core Intel i7, and mounting an embedded Nvidia GeForce GTX 1050. Hyperparameters, such as number of epochs and batch size, were chosen by applying the early-stopping method to the validation sets. For the detection phase, OpenPose model is based on COCO [45], and is designed to provide up to 25 body landmarks, 70 face landmarks, 42 hands landmarks and 6 feet landmarks for each target. However, for the purpose of this work, only 14 body landmarks were exploited. Specifically, since 5 out of 25 body landmarks represent *nose*, *left eye*, *right eye* and *left* and *right ears*, we defined a *head* landmark averaging these landmarks. Thus, in our implementation, the set of body landmarks is $J = \{1, \ldots, 14\}$, as described in Figure 1. Finally, regarding ActionXPose coding, feature computations were conducted in MATLAB, while the classification was performed using the Keras implementation of MLSTM-FCN, provided by [14].

D. Results

In this section, we provide results for ActionXPose performance on the three set of features defined in Section III-E, i.e. Spatial-attention, Temporal-attention and Spatio-temporal-attention. Simulations were conducted for the three experimental settings defined in Section IV-B. Obtained results are provided in Table III and compared with the state-of-the-art baselines methods. Regarding the Traditional experimental setting, the action classes are unbalanced due to the presence

TABLE III
ACCURACY RESULTS (%) FOR THREE EXPERIMENTAL SETTINGS, I.E. TRADITIONAL (TRAD.), DATASET AUGMENTATION (D. AUGM.) AND CROSS-DATASET (CROSS-D.). $|\mathcal{L}|$ REPRESENTS THE NUMBER OF CONSIDERED ACTIONS.

Settings	**Trad.**	**D. Augm.**	**Cross-D.**
$\|\mathcal{L}\|$	**18 / 17**	**18 / 17**	**17**
[19] + [14]	92.44 / 88.88	93.77 / 91.07	93.77
[18] + [14]	81.77 / 73.70	79.55 / 75.00	79.55
[18] + [19] + [14]	91.99 / 87.96	84.00 / 80.73	84.00
Spatial-attention	91.55 / 86.24	96.00 / 96.33	96.00
Temporal-attention	94.22 / 91.74	92.88 / 89.29	92.88
Spatio-temporal attention	95.11 / 92.73	96.44 / 95.58	96.44

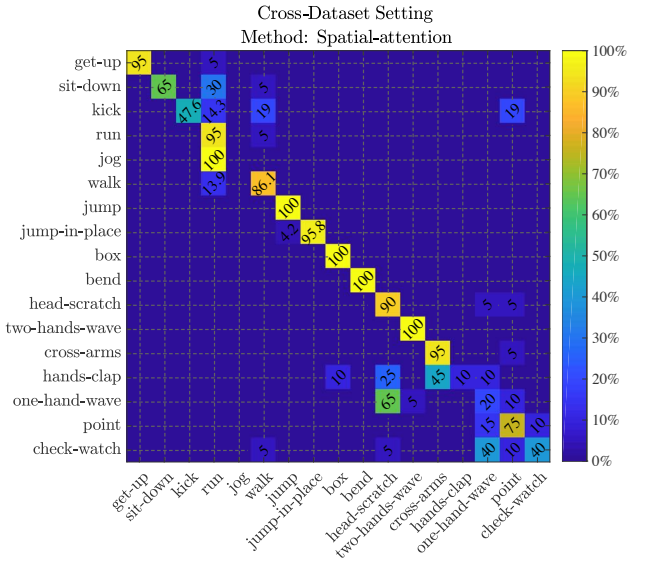


Fig. 10. Confusion matrix obtained for the Spatial-attention method in the Cross-Dataset Setting.

of the *stand* action. Thus, results for this setting were normalized using the total number of clips per action.

Overall, ActionXPose features outperform the baselines in almost all tests. Moreover, in the Dataset Augmentation experimental setting, additional training data improves the results obtained in the Traditional experimental setting. This is mainly due to the higher generalization degree obtained by providing additional MPOSE data during the training phase. In the Cross-Dataset setting, obtained results shows that MPOSE does not contain enough data variability to fully meet the ISLD requirements. However, it surprisingly covers most of the actions, confirming to be a good pre-training source of data. In Fig. 10, the confusion matrix obtained in the Cross-Dataset setting is provided.

E. Ablation Study

In this section, different combinations of low and high-level features are considered. This ablation study was conducted on the MPOSE dataset. We adopted the cross-validation setting used in many different works [46]–[53]. Specifically, ActionXPose methods were tested by using an action-based cross-validation setting with 10 foldings. This approach stabilizes the number of samples per action across different foldings. Pose-

TABLE IV

ACTIONXPOSE ABLATION STUDY BASED ON MPOSE DATASET. (LEFT) PERFORMANCE OF DIFFERENT METHODS IS PROVIDED ON AVERAGE (AVG) FOR 10 CROSS-VALIDATION FOLDINGS, REPORTING OBTAINED STANDARD DEVIATIONS (STD). THE **Feat** COLUMN REPORTS THE ACTUAL NUMBER OF FEATURES AVAILABLE FOR EACH METHOD. REGARDING HYPER-PARAMETERS, DEFAULT AND $\sigma = 2, z = 2$ CASES ARE REPORTED. (RIGHT) PAIRED T-TEST P-VALUES WITH $\alpha = 0.05$ ARE REPORTED. P-VALUES WHERE SIGNIFICANT DIFFERENCES BETWEEN METHODS ARE NOT SUPPORTED BY THE CONDUCTED TESTS (P-VALUE $> \alpha$) ARE HIGHLIGHTED IN BOLD.

		Default		$\sigma = 0.2, z = 2$	
Method	**Feat**	**AVG (%)**	**STD**	**AVG (%)**	**STD**
A	28	91.47	1.68	93.87	1.26
B	28	91.72	1.20	92.64	1.43
C	17	88.90	1.50	91.88	1.68
D	68	91.31	2.03	92.11	1.44
ABCD	141	92.95	0.89	95.48	1.34
F	17	77.53	3.29	78.47	5.77
G	68	84.28	2.36	86.60	2.05
ABFG	141	93.69	1.61	95.43	1.14
all	226	92.54	1.87	94.44	0.81

Significance for $\sigma = 0.2, z = 2$								
	B	**C**	**D**	**ABCD**	**F**	**G**	**ABFG**	**all**
A	.025	.009	.000	.000	.000	.000	.003	.009
B	-	**.139**	**.211**	.000	.000	.000	.000	.001
C	-	-	**.690**	.000	.000	.000	.000	.000
D	-	-	-	.000	.000	.000	.000	.000
ABCD	-	-	-	-	.000	.000	**.919**	.030
F	-	-	-	-	-	.001	.000	.000
G	-	-	-	-	-	-	.000	.000
ABFG	-	-	-	-	-	-	-	.001

flipping and pose-noising were applied to training samples, while validation and testing samples were not augmented.

The following ablated methods were considered:

$$\begin{aligned}
&\textbf{A} && p(t) \\
&\textbf{B} && p'(t) \\
&\textbf{C} && D_{V_l,J}(t),\ \forall l \in \mathcal{L} \\
&\textbf{D} && [D_{V_l,J_a}(t),\ \ldots,\ D_{V_l,J_d}(t)],\ \forall l \in \mathcal{L} \\
&\textbf{ABCD} && [p(t),\ p'(t),\ Seq(V_l)],\ \forall l \in \mathcal{L} \\
&\textbf{F} && D_{S_l,J}(t),\ \forall l \in \mathcal{L} \\
&\textbf{G} && [D_{S_l,J_a}(t),\ \ldots,\ D_{S_l,J_d}(t)],\ \forall l \in \mathcal{L} \\
&\textbf{ABFG} && [p(t),\ p'(t),\ Seq(S_l)],\ \forall l \in \mathcal{L} \\
&\textbf{all} && [p(t),\ p'(t),\ Seq(V_l),\ Seq(S_l)],\ \forall l \in \mathcal{L}
\end{aligned} \tag{20}$$

where the i-subscript has been omitted for convenience. We remark that **ABCD** and **ABFG** correspond to the Spatial-attention and Temporal-attention methods, respectively, defined in Section III-E. Similarly, **all** corresponds to the method Spatio-temporal-attention method. This new nomenclature better highlights the method compositions in terms of features.

Averaged results for these methods are shown in Table IV, including standard deviations, for the Default and $\sigma = 0.2, z = 2$ cases. It turns out that the latter case the performance is superior due to the data augmentation provided by pose-flipping and pose-noising.

In order to measure the significance of the obtained performance, paired t-tests with $\alpha = 0.05$ were conducted for the $\sigma = 0.2, z = 2$ case. Since we expect that the more features involved in the learning process, the higher the averaged accuracy, a one-tail paired t-test was chosen whenever possible. In all other cases, i.e. when the number of features were the same for both compared methods, a two-tail paired t-test was chosen. We report p-values for all paired methods in Table IV.

Overall, this ablation study highlights that the ActionXPose high-level features always bring benefits to the learning process when considered alongside low-level features, i.e. $p(t)$ and $p'(t)$. Moreover, the **all** method is not the best for MPOSE. This shows that, when too many features are involved, it is difficult for the network to effectively extract useful information. Therefore, trade-off methods are more effective. Indeed, **ABCD** and **ABFG** are significantly the best methods. This could be possibly due to the *curse of dimensionality* which occur when the number of features is too high with respect the number of available training samples [54]. Therefore, this study also highlights the importance of data augmentation, which considerably improves the *Default* case performance.

Overall, as expected, this test shows the importance of the low-level features for carrying most of the action knowledge. In fact, methods **A** and **B** achieve very good performance. On the other hand, as we saw in Section IV-F, these methods are strongly affected by occlusions. Therefore, high-level based methods provide the required robustness, along with clear advantages in terms of accuracy.

F. Occlusions Study

In this section, the robustness of ActionXPose features to occlusions and missing data is evaluated. In particular, the strategies proposed in sections III-D1, III-D2, and III-D4 for occlusion-handling will be compared. The short-time occlusions strategy in Section III-D3 is always applied. Moreover, in all experiments, pose-flipping and pose-noising were always applied to training data, with $\sigma = 0.2$ and $z = 1$. The experiment is based on MPOSE dataset. Since MPOSE data contains only self-occlusions, we simulated more challenging occlusions by explicitly removing landmarks from the testing data. This strategy is fast and effective, does not require any time-consuming video editing, and provides similar results as assumed the occlusions are in the video data. Inspired by landmark subsets in (12), we removed 6 different groups of landmarks, i.e.

$$\begin{aligned}
J_a^* &= \{4, 5\} && \text{(Right Arm)} \\
J_b^* &= \{7, 8\} && \text{(Left Arm)} \\
J_c^* &= \{10, 11\} && \text{(Right Leg)} \\
J_d^* &= \{13, 14\} && \text{(Left Leg)} \\
J_{a,b}^* &= \{4, 5, 7, 8\} && \text{(Both Arms)} \\
J_{c,d}^* &= \{10, 11, 13, 14\} && \text{(Both Legs)}
\end{aligned} \tag{21}$$

It is worth emphasizing that the baseline methods are strongly numerically affected by the proposed occlusions. In other words, such occlusions create persistent missing data, and thus persistent missing features. On the other hand, the

ActionXPose high-level features are more numerically robust, due to the definition of the embedding distance in (13). Specifically, when such occlusions occur, the proposed high-level features only slightly change their values, rather than being completely lost.

The first experiment (Fig. 11-Top) consists of occluding testing data, without performing neither occlusions augmentation nor landmarks borrowing techniques. Thus, the trained networks were not prepared to face such occlusions. As expected, the baseline methods are strongly less robust than ActionXPose features. In contrast, all methods that include high-level features achieve much better performance due to the robustness provided by equations (13) and (14). In particular, the proposed Spatio-temporal attention method remarkably outperforms the baselines in all the occlusion cases.

In the second experiment (Fig. 11-Middle), we enabled occlusions augmentation only and repeated the same experiment mentioned above. In this case, since the training data include synthetically occluded data, the resulting networks are much more robust to occlusions. In this case, baseline methods are also expected to be more robust since the trained network is prepared to deal with the missing features carried by low-level sequences. However, again, high-level features outperform the baselines in all cases.

In the third and last experiment (Fig. 11-Bottom), we enabled both occlusions augmentation and landmarks borrowing. The borrowing landmark technique is able to fill the gaps due to occlusions in the baseline features. To perform this experiment, training and validation data were firstly occluded by the occlusions augmentation technique, while testing data were occluded with the proposed equation (21). Then, all low-level and high-level features for training, validation and testing data were computed considering the borrowing landmarks technique. The first effect of this processing is that performance and robustness globally further increase. However, again, the proposed ActionXPose features outperform the baselines in all occlusion cases.

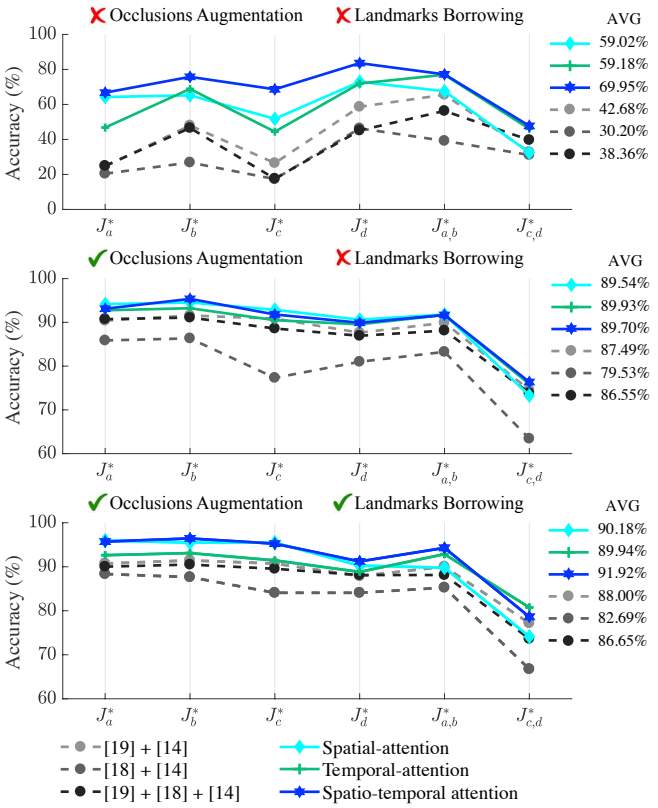


Fig. 11. Occlusion Study results, provided on average over 10 cross-validation foldings. AVG represents the averaged results over the six occlusion cases. (Top) Performance with no data-augmentation nor borrowing landmarks. (Middle) Performance with data-augmentation but no borrowing landmarks. (Bottom) Performance with data-augmentation and borrowing landmarks.

G. Performance on Traditional Datasets

In this section, ActionXPose results on the KTH and i3DPost datasets are provided. These tests were conducted to allow comparisons between the proposed method and other state-of-the-art methods. Since KTH and i3DPost include specific challenges, such as multiple viewpoints, zooming in/out, moving cameras, and variable target-camera proximity, this test can also show ActionXPose robustness against these challenges.

Tests are performed on the KTH dataset under two experimental settings. The first is the Split setting, where training, validation and testing samples are predefined by the original author in [41]. The second is the Leave-One-Actor-Out (LOAO) setting, where multiple tests are conducted by using each actor as testing actor and averaging obtained results. Table V shows the results for both these experimental settings. Data augmentation parameters were empirically chosen and fixed for all tests as $z = 0$ and $\sigma = 0.2$.

The ActionXPose high-level features outperform the baseline methods in all settings. Moreover, in the Split setting, ActionXPose performance is among the state-of-the-art. In the case of the LOAO setting, ActionXPose outperforms other state-of-the-art methods.

Regarding i3DPost, this dataset is usually tested under the LOAO setting. i3DPost is specifically designed to cover multi-viewpoints HAR. In fact, it includes video clips recorded from 8 different viewpoints. The results are given in the most difficult case where the task was to recognize the target action by using data from one single camera in a multi-viewpoint mode. ActionXPose results are summarized in Table V and compared with the state-of-the-art. As in the previous test, ActionXPose outperforms the baseline methods, achieving state-of-the-art performance.

In conclusion, these tests are particularly suitable for highlighting the effectiveness of pose-based HAR in comparison with the traditional methods. In fact, such excellent results were obtained by using 2D human poses only, while the other state-of-the-art methods exploited RGB data or other sophisticated data sources such as human silhouette. In this work, we showed that 2D pose-based HAR can achieve similar, and sometimes superior, performance than traditional RGB based methods.

H. Computational Speed Evaluation

In this section, we provide computational speed evaluations for each of the most important step required by ActionXPose.

TABLE V
ACCURACY RESULTS (%) FOR KTH AND I3DPOST. FOR KTH DATASET, TWO SETTINGS ARE REPORTED, NAMELY SPLIT AND LOAO. FOR I3DPOST DATASET, LOAO RESULTS ARE OBTAINED FOR $|\mathcal{W}| = 8$ AND $|\mathcal{W}| = 2$ (*), WHERE $|\mathcal{W}|$ REPRESENTS THE NUMBER OF CONSIDERED VIEWPOINTS.

	KTH		i3DPost
Method	**Split**	**LOAO**	**LOAO**
Spatial-attention	90.50	**99.04**	98.95
Temporal-attention	90.15	98.03	98.95
Spatio-temporal attention	89.80	98.26	99.47
[19] + [14]	88.06	98.91	97.39
[18] + [14]	83.19	96.29	95.30
[19] + [18] + [14]	86.44	96.67	99.47
Kovashka et al. [26]	**94.50**	-	-
Zhang et al. [27]	94.10	-	-
Ji et al. [28]	90.20	-	-
Almeida et al. [29]	-	98.00	-
Vrigkas et al. [30]	-	98.30	-
Liu et al. [31]	-	93.80	-
Raptis and Soatto [32]	-	94.50	-
Jiang et al. [33]	-	95.77	-
Gilbert et al. [34]	-	95.70	-
Angelini et. al. [10]	-	-	**99.73**
Castro et al. [46]	-	-	99.00(*)
Iosifidis et al. [47]	-	-	98.16
Azary et al. [48]	-	-	92.97
Hilsenbeck et al. [49]	-	-	92.42

We anticipate that, our simulations showed that the body pose detector is the *bottleneck* for the entire processing. However, it is claimed to be a real-time detector [3] when hardware requirements are fully satisfied. This statement is supported by our experiments.

This computational speed evaluation considerers *training* and *testing* phases separately.

1) Training phase: in the training phase, the most intense step (excluding the body pose detection) is the library creation step, where the SOM clustering method is performed. The SOM computational complexity can be estimated as $\mathcal{O}(N^2C^2)$ [55], where N is the number of considered samples and C is the vector input dimensionality. This result shows the importance of considering the PCA as a dimensionality reduction technique before running the SOM. Another important step is sequence embedding step, which has a complexity of $\mathcal{O}(n)$, where n is the number of prototypes in the considered library [10]. Thus, for a small n, such as in our case, the embedding process is remarkably fast. Our simulations showed that the embedding step can process a single frame in $\approx 9 * 10^{-5}$ seconds, which corresponds to around 10^4 frame per second. Even considering that ActionXPose needs to run two embedding processes (for the spatial and temporal libraries), the embedding process is still very fast.

2) Testing phase: in the testing phase, the most computationally intense step (excluding the body pose detection) is the sequence embedding. For the **all** method, each clip sequence requires on average $4.1 * 10^{-2}$ seconds to be processed. Since in our tests each clip contains on average 72 frames, ActionXPose can elaborate its prediction with a speed of $\approx 5.7 * 10^{-4}$ seconds per frame. Thus, once hardware requirements for the OpenPose detector are satisfied, ActionXPose can run with real-time performance.

TABLE VI
VIDEO QUALITY COMPARISON, REPORTING FOR EACH USED DATASET, COLOUR CHANNELS (CHAN.), FRAME-PER-SECONDS (FPS), MEGA-BITS-PER-SECOND (MBITS/S), FRAME SIZE AND AVERAGED BODY SIZE.

Dataset	**Chan.**	**FPS**	**mbits/s**	**Frame Size**	**Body Size**
Weizmann	RGB	25	15.55	180x144	65x93
i3DPost	RGB	25	5.18	960x540	384x408
IXMAS	RGB	19	1.9	390x291	136x73
KTH	mono	25	0.89	160x120	82x106
ISLD	RGB	25	47.97	1920x1080	403x557

I. Varying Video Quality Study

In this section, we provide additional insights about the proposed method robustness to different frame resolutions, colour channels, frame-per-second (FPS), mega-bits-per-second (mbits/s), actual body size and frame quality rate. In fact, it is reasonable to expect that different video qualities, in terms of the above-mentioned indicators, might result in different OpenPose performance, which in turn can affect ActionXPose performance. In Table VI, the datasets used in this work are compared in terms of these common video indicators. The variety of conditions shown in Table VI, compared with the performance presented in previous sections, demonstrates that ActionXPose performance is stable across different conditions.

As an additional study, we conducted further experiments on ISLD, under the Dataset Augmentation Setting presented in Section IV-B2. The first goal was to assess the impact of varying frame sizes and body sizes on ActionXPose performance. To this purpose, we repeatedly reduced the original ISLD frame size by a factor of 5. Each time, the resulting clips were saved in AVI format with the Motion JPEG 2000 encoder provided by MATLAB, with a quality threshold of 95%. We report obtained results in Figure 12-Top. The second goal was to assess the impact of varying Motion JPEG quality rates on ActionXPose performance. Therefore, we set the frame size to a reasonable value, i.e. 192x108 px, and then repeatedly reduced the quality threshold from 95% to 35%. We report obtained results in Figure 12-Bottom.

Overall, conducted tests show that body size reduction is slightly related to a reduction of OpenPose performance. In contrast, ActionXPose performance remains quite stable. Surprisingly, ActionXPose performance peak is not achieved on the best frame size. In fact, the peak of performance is expected to be in correspondence with the most similar training condition. Similarly, the quality rate slightly worsens OpenPose performance. This further false negative increment seems to have a limited impact on ActionXPose performance.

In conclusion, these results suggest that ActionXPose is robust to the studied working conditions. However, further improvements could be performed by introducing a time-related dropout layer in the network, to impose even more robustness to sudden false negative detections. However, this is going to be part of future studies.

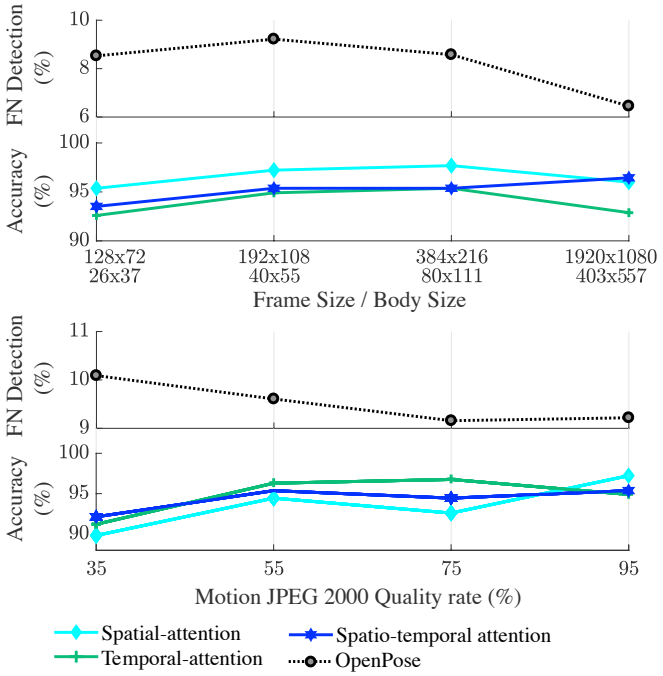


Fig. 12. Frame size, body size and quality rate impact on OpenPose in terms of False Negative Detection rate (FN), compared with ActionXPose performance in terms of accuracy. (Top) Impact of the frame size and body size changes on OpenPose and ActionXPose methods. (Bottom) Impact of the frame quality changes on OpenPose and ActionXPose methods.

J. Discussions

Overall, as shown in Tables III and V, the proposed features outperform the baselines methods in almost all tests. In particular, in the case of KTH dataset (LOAO), the proposed Spatial-Attention method outperforms the state-of-the-art. Regarding table V, state-of-the-art methods exploit raw RGB data, human silhouettes and motion history maps as main learning sources. These sources contain comprehensive and redundant information, which are expected to be more informative than poses only. However, we showed that by using *limited data*, i.e. 2D human poses with 14 landmarks and no other pixel-based information, it is possible to achieve state-of-the-art performance. The advantages of using poses rather than pixel-based approaches are threefold. First, it turns out to be light processing, which can be performed right after target detection, allowing native integration with existing tracking systems. Second, it is a self-explaining method: poses can be visually checked by human operators and compared with the predicted action labels, for an effective troubleshooting on-the-site. Third, poses are human-appearance and context insensitive, which allow straightforward cross-datasets implementation.

As shown in Section IV-H, the proposed method is remarkably fast in terms of computational speed. Thus, ActionXPose fits into a tracking system without additional, expensive, computational costs.

Moreover, the major advantage of ActionXPose is in occlusions and missing data challenges. Since OpenPose is relatively prone to false detections, a robust method was required for pose-based HAR. As shown in Fig. 11, the baseline methods are not suitable to face body limbs occlusions. In contrast, the proposed Spatio-temporal attention method shows greater robustness when body limbs occlusions occur.

Despite the advantages mentioned above, our experiments also showed OpenPose limitations. In fact, in the case of UCF101 and HMDB51 datasets, OpenPose's low performance was one of the causes that compromised the proposed processing. As shown in Table VI and Figure 12, frame resolution and frame quality are not a major problem themselves. In contrast, as the UCF101 and HMDB51 results suggest, the major issues were due to strong ambiguity between the target and the background. Moreover, the very small ratio between the target size and the frame resolution also compromised the detection.

The above mentioned OpenPose limitations have an impact on ActionXPose. In fact, when the link vectors $v_{2,9}$ and $v_{2,12}$ are both persistently missing, the strategy in Section III-B is not well-posed. However, this case occurs only when the body trunk is persistently occluded or undetectable. In this case, even human eyes might fail in classifying posture-related actions such as those studied in this work. Such disruptive cases require ad-hoc studies which are beyond the purpose of this work.

Another problem occurs when more challenging action classes are considered for HAR, for example, those provided by UCF101 and HMDB51 datasets. Since the UCF101 and HMDB51 are *not* posture-related datasets, colour contextual information is crucial, and the poses are not informative enough to fully describe the performed action. Further studies can be conducted to consider *colors* alongside *poses*, to combine both approaches in a multimodal system. However, this solution is beyond the purpose of this work. In fact, our goal was to provide a posture-related method for HAR that was able to fit the strict computational requirements of a security system. Thus, in this work, we showed that exploiting the OposePose detection, which is already used for tracking, we can perform robust pose-based HAR without expensive additional costs.

V. Conclusions and Future Work

In this work, we presented the ActionXPose algorithm for 2D pose-based HAR, which achieves state-of-the art performance on selected datasets. Proposed high-level features improve accuracy and robustness to occlusions and missing data in comparison with the low-level features based method. In addition, this work proposed a new dataset for pose-level HAR in CCTV-like environment, namely ISLD dataset. This datasets was used to extensively test several variations of the proposed method, under different experimental conditions, including the interesting Dataset Augmentation and Cross-Dataset settings.

Future work will mainly focus on three directions. First, more generalization ability is required to allow the system to work in more complex scenarios, without additional training. Second, RGB based processing could be helpful for considering target appearance-related information. However, since privacy issues and computational speed have to be considered

for security and surveillance applications, tailored solutions are required. Third, ActionXPose can be effectively made ready-to-use in surveillance scenarios once an automatic *action detection* for online surveillance video sequences will be integrated. However, such abilities requires ad-hoc solutions and further studies.

ACKNOWLEDGMENT

The authors would like to thank Thales and EPSRC for supporting this project with the Industrial Cooperative Awards in Science & Technology (CASE).

REFERENCES

[1] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and Vision Computing*, vol. 60, pp. 4–21, 2017.

[2] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A Review on Human Activity Recognition Using Vision-Based Method," *Journal of Healthcare Engineering*, vol. 2017, pp. 1–31, 2017.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," in *Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.

[4] P. Feng, W. Wang, S. M. Naqvi, and J. Chambers, "Adaptive Retrodiction Particle PHD Filter for Multiple Human Tracking," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1592–1596, 2016.

[5] A. Ur-Rehman, S. M. Naqvi, L. Mihaylova, and J. A. Chambers, "Multi-Target Tracking and Occlusion Handling With Learned Variational Bayesian Clusters and a Social Force Model," *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1320–1335, 2016.

[6] P. Feng, W. Wang, S. Dlay, S. M. Naqvi, and J. Chambers, "Social Force Model-Based MCMC-OCSVM Particle PHD Filter for Multiple Human Tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 725–739, 2017.

[7] Z. Fu, P. Feng, F. Angelini, J. Chambers, and S. M. Naqvi, "Particle PHD Filter Based Multiple Human Tracking Using Online Group-Structured Dictionary Learning," *IEEE Access*, vol. 6, pp. 14 764–14 778, 2018.

[8] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi, "Multi-Level Cooperative Fusion of GM-PHD Filters for Online Multiple Human Tracking," *IEEE Transactions on Multimedia*, p. 1, 2019.

[9] O. Barnich and M. V. Droogenbroeck, "ViBe : A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[10] F. Angelini, Z. Fu, S. A. Velastin, J. A. Chambers, and S. M. Naqvi, "3D-Hog Embedding Frameworks for Single and Multi-Viewpoints Action Recognition Based on Human Silhouettes," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4219–4223.

[11] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

[12] P. Wang and P. O. Ogunbona, "RGB-D-based Motion Recognition with Deep Learning: A Survey," *International Journal of Computer Vision (IJCV)*, vol. TBA, no. June, pp. 1–34, 2017.

[13] T. Kohonen, *Self-organizing maps*, 3rd ed. Berlin: Springer, 2001.

[14] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for Time Series Classification," *arXiv*, 2018.

[15] S. Hochreiter and J. Urgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] B. Liang and L. Zheng, "A Survey on Human Action Recognition Using Depth Sensors," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2015, pp. 1–8.

[17] B. Langmann, K. Hartmann, and O. Loffeld, "Depth Camera Technology Comparison and Performance Evaluation," *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, vol. 2, pp. 438–444, 2012.

[18] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1290–1297.

[19] A. Taha, H. H. Zayed, and M. E. Khalifa, "Skeleton-based Human Activity Recognition for Video Surveillance," *International Journal of Scientific & Engineering Research*, vol. 6, no. 1, pp. 993–1004, 2015.

[20] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Lecture Notes in Computer Science*, vol. 9910 LNCS, 2016, pp. 34–50.

[21] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," *arXiv*, 2018.

[22] J. Liu, A. Shahroudy, D. Xu, A. Kot Chichung, and G. Wang, "Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[23] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 07-12-June, 2015, pp. 1110–1118.

[24] V. Veeriah, Z. Naifan, and G.-J. Qi, "Differential Recurrent Neural Networks for Action Recognition," in *International Conference on Computer Vision (ICCV)*, 2015, pp. 4041–4049.

[25] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks," in *Conference on Artificial Intelligence (AAAI)*, 2016, pp. 3697–3703.

[26] A. Kovashka and K. Grauman, "Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition," *Computer Vision and Pattern Recognition (CVPR)*, pp. 2046–2053, 2010.

[27] Y. Zhang, X. Liu, M. C. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 707–721.

[28] S. Ji, M. Yang, K. Yu, and W. Xu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–31, 2013.

[29] R. Almeida, Z. K. Goncalves Do Patrocinio, and S. J. F. Guimaraes, "Exploring quantization error to improve human action classification," in *International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 1354–1360.

[30] M. Vrigkas, V. Karavasilis, C. Nikou, and I. A. Kakadiaris, "Matching mixtures of curves for human action recognition," *Computer Vision and Image Understanding*, vol. 119, pp. 27–40, 2014.

[31] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1996–2003.

[32] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 577–590.

[33] Z. Jiang, Z. Lin, and L. Davis, "Recognizing actions by shape-motion prototype trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 533–547, 2012.

[34] A. Gilbert, J. Illingworth, and R. Bowden, "Action Recognition Using Mined Hierarchical Compound Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 883–897, 2011.

[35] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[36] H. Jo, Y. H. Na, and J. B. Song, "Data augmentation using synthesized images for object detection," in *International Conference on Control, Automation and Robotics (ICCAR)*, 2017, pp. 1035–1038.

[37] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised Domain Adaptation for Robust Speech Recognition via Variational Autoencoder-Based Data Augmentation," *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 16–23, 2017.

[38] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *arXiv*, 2012.

[39] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[40] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *arXiv*, 2016.

[41] C. Schuldt, L. Barbara, and S. Stockholm, "Recognizing Human Actions: A Local SVM Approach," *International Conference on Pattern Recognition (ICPR)*, vol. 3, pp. 32–36, 2004.

[42] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, pp. 249–257, 2006.

[43] R. B. L. Gorelick, M. Blank, E. Shechtman, M. Irani, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[44] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3DPost multi-view and 3D human action/interaction database," in *European Conference for Visual Media Production (CVMP)*, 2009, pp. 159–168.

[45] T.-Y. Lin, C. L. Zitnick, and P. Doll, "Microsoft COCO: Common Objects in Context," *arXiv*, 2015.

[46] G. Castro-Munoz and J. Martinez-Carballido, "Real Time Human Action Recognition Using Full and Ultra High Definition Video," in *International Conference on Computational Science and Computational Intelligence (CSCI)*, 2015, pp. 509–514.

[47] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis," *Signal Processing*, vol. 93, no. 6, pp. 1445–1457, 2013.

[48] S. Azary and A. Savakis, "Multi-view action classification using sparse representations on Motion History Images," in *Western New York Image Processing Workshop (WNYISPW)*, 2012, pp. 5–8.

[49] B. Hilsenbeck, D. Munch, H. Kieritz, W. Hubner, and M. Arens, "Hierarchical Hough forests for view-independent action recognition," in *International Conference on Pattern Recognition (ICPR)*, 2016, pp. 1911–1916.

[50] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[51] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *International Conference on Computer Vision (ICCV)*, vol. II, 2005, pp. 1458–1465.

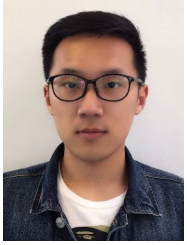
[52] A. Berg, T. Berg, and J. Malik, "Shape Matching and Object Recognition Using Low Distortion Correspondences," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 26–33.

[53] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer-Verlag New York, 2013.

[54] I. Goodfellow, *Deep learning*. The MIT Press, 2016.

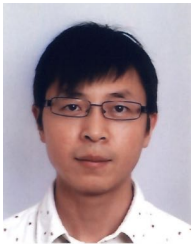
[55] D. Roussinov and H. Chen, "A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation," *Communication Cognition and Artificial Intelligence (CC-AI)*, vol. 15, no. 1-2, pp. 81–111, 1998.

Federico Angelini (S'17) received the Master degree in Pure and Applied Mathematics from University of Rome "Tor Vergata", Italy, in 2015. He collaborated with the National Research Council (CNR), Rome, Italy on the Image Denoising Methods project in 2015-2016. He is currently pursuing the PhD degree within the Intelligent Sensing and Communications (ISC) Research Group, School of Engineering, Newcastle University, UK. His PhD project is funded by the iCase Award by EPSRC/Thales, UK. His research interests include Human Action and Behaviour Recognition via Multimodal Sensors, Machine Learning, Deep Learning and Signal Processing.



Zeyu Fu (S'15-M'18) received the B.Eng. degree in electrical and electronic engineering (with First-Class Hons.) from Newcastle University, Newcastle Upon Tyne, U.K., in 2015, where he is at the final stage of his Ph.D. degree within Intelligent Sensing and Communications Research Group, School of Engineering in 2019. He is currently working as a postdoc in the same group in Newcastle University. His research interests include video tracking, machine learning, and medical image analysis.

Yang Long is an Assistant Professor in the Department of Computer Science, Durham University. He is also an MRC Innovation Fellow aiming to design scalable AI solutions for large-scale healthcare applications. His research background is in the highly interdisciplinary field of Computer Vision and Machine Learning. While he is passionate about unveiling the black-box of AI brain and transferring the knowledge to seek Scalable, Interactable, Interpretable, and sustainable solutions for other disciplinary researches, e.g. physical activity, mental health, design, education, security, and geoengineering. He has authored/co-authored 30+ top-tier papers in refereed journals/conferences such as IEEE TPAMI, TIP, CVPR, AAAI, and ACM MM, and holds a patent and a Chinese National Grant.



Ling Shao is the CEO and Chief Scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include computer vision, machine learning, and medical imaging. He is a fellow of IAPR, IET, and BCS. He is an Associate Editor of the IEEE Transactions on Neural Networks and Learning Systems, and several other journals.

Syed Mohsen Naqvi (S'07-M'09-SM'14) received the Ph.D. degree in Signal Processing from Loughborough University, Loughborough, U.K., in 2009 and his Ph.D. thesis was on the EPSRC U.K. funded project. He was a Postdoctoral Research Associate on the EPSRC U.K.-funded projects and REF Lecturer from 2009 to 2015. Prior to his postgraduate studies in Cardiff and Loughborough Universities U.K., he served the National Engineering and Scientific Commission (NESCOM) of Pakistan from 2002 to 2005.

Dr Naqvi is Lecturer/Assistant Professor in Signal and Information Processing at the School of Engineering, Newcastle University, Newcastle, U.K. His research interests include multimodal processing for human behaviour analysis, multi-target tracking, and source separation all; for machine learning. He organized special sessions in FUSION, delivered seminars and was a speaker at UDRC Summer Schools 2015-2017. He has 100+ publications with the main focus of his research being on Multimodal (audio-video) Signal and Information Processing. He is an Associate Editor for Elsevier Journal on Signal Processing. He is Fellow of the Higher Education Academy (FHEA). He is an Associate Editor for IEEE Transactions on Signal Processing.